THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 48
Standardized Variables

Chris A. Mack
Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Standardizing Variables

- Consider the model

$$y = f(\boldsymbol{x}, \boldsymbol{\beta}) + \varepsilon \qquad \boldsymbol{x} = (x_1, x_2, \ldots); \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots)$$

- We can "standardize" each of the variables

	response	jth predictor
Standardized (unit normal scaling)	$\tilde{y}_i = \frac{y_i - \bar{y}}{s_y}$	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$
Correlation Transformation (unit length scaling)	$\tilde{y}_i = \frac{y_i - \bar{y}}{\sqrt{n-1}s_y}$	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{n-1}s_j}$

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Recall OLS Matrix Formulation

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{bmatrix} = design\ matrix$$

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

(each row in X and Y is a "data point")

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Recall OLS Matrix Formulation

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{HY}$$

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = hat\ matrix$$

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

$$cov(\hat{\boldsymbol{\beta}}) = s_e^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Standardized Matrix Formulation

- The "standardized regression model" uses the correlation transformation
  - Note: it will not have an intercept term β_0

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon} \qquad \tilde{\boldsymbol{X}} = \begin{bmatrix} \tilde{x}_{11} & \cdots & \tilde{x}_{1,p-1} \\ \vdots & \ddots & \tilde{x}_{2,p-1} \\ \tilde{x}_{n1} & \cdots & \tilde{x}_{n,p-1} \end{bmatrix} = design\ matrix$$

Converting back: $\boldsymbol{\beta}_j = \tilde{\boldsymbol{\beta}}_j\left(\frac{s_y}{s_j}\right) \qquad \boldsymbol{\beta}_0 = \bar{y} - \sum_{j=1}^{p-1}\boldsymbol{\beta}_j\bar{x}_j$

$$SE(\boldsymbol{\beta}_j) = SE(\tilde{\boldsymbol{\beta}}_j)\left(\frac{s_y}{s_j}\right)$$

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Standardized Matrix Formulation

- The $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}$ are correlation matrices

$$\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1,p-1} \\ r_{12} & 1 & \cdots & r_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1n} & r_{2n} & \cdots & 1 \end{bmatrix} \qquad \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}} = \begin{bmatrix} r_{1y} \\ \vdots \\ r_{ny} \end{bmatrix}$$

- where r_{ij} is the correlation between the ith and jth regressors, r_{iy} is the correlation between the ith regressor and y

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Consider a Two Regressor Model

$$\tilde{y} = \tilde{\beta}_1 \tilde{x}_1 + \tilde{\beta}_2 \tilde{x}_2 + \varepsilon \qquad (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1} = \begin{bmatrix} \frac{1}{(1-r_{12}^2)} & \frac{-r_{12}}{(1-r_{12}^2)} \\ \frac{-r_{12}}{(1-r_{12}^2)} & \frac{1}{(1-r_{12}^2)} \end{bmatrix}$$

$$R^2 = \frac{r_{1y}^2 + r_{2y}^2 - 2r_{12}r_{1y}r_{2y}}{(1-r_{12}^2)} = \tilde{\beta}_1 r_{1y} + \tilde{\beta}_2 r_{2y} \qquad cov(\tilde{\beta}_1, \tilde{\beta}_2) = \frac{-r_{12}s_\varepsilon^2}{(1-r_{12}^2)}$$

$$\tilde{\beta}_1 = \frac{r_{1y} - r_{12}r_{2y}}{(1-r_{12}^2)} \qquad \tilde{\beta}_2 = \frac{r_{2y} - r_{12}r_{1y}}{(1-r_{12}^2)} \qquad SE(\tilde{\beta}_1) = SE(\tilde{\beta}_2) = \frac{s_\varepsilon}{\sqrt{1-r_{12}^2}}$$

$$\tilde{h}_{ii} = h_{ii} = \frac{1}{n} + \frac{1}{(1-r_{12}^2)}\left(\tilde{x}_{1i}^2 + \tilde{x}_{2i}^2 - 2r_{12}\tilde{x}_{1i}\tilde{x}_{2i}\right)$$

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Effects of Standardization

- Standardizing the variables can sometimes be useful
  - Reduces correlations between $\tilde{x}$ and $\tilde{x}^2$, and between interactions and primary terms
  - Less numerical instabilities for matrix inversion and OLS solution
- Standardization will not help with multicollinearity
  - We will use standardization later, with principle component analysis (PCA)

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Lecture 48: What have we learned?

- What is variable standardization, and why is it used?
- In general, will standardization help with problems of multicollinearity?